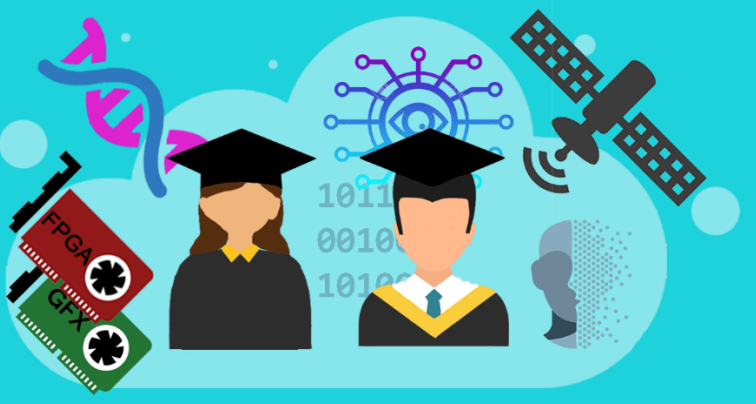
# Diploma Thesis

Microproccessors and Digital Systems Laboratory




## Modeling NVIDIA Multi-Instance GPU Devices

Graphics Processing Units (GPUs) have revolutionized computing by enabling high-performance acceleration for a wide range of general-purpose applications. Initially designed for rendering graphics, GPUs have become indispensable in accelerating compute-intensive tasks, especially in the field of deep neural networks (DNNs). With the explosion of artificial intelligence (AI) and machine learning (ML) applications, the demand for GPU resources has surged, pushing the boundaries of what these devices can achieve.

In response to the growing demand for GPU resources, modern GPU architectures have introduced multi-instance capabilities. Multi-Instance GPU (MIG) technology enables the partitioning of a single physical GPU into multiple smaller, isolated instances, each capable of running separate workloads. This innovation improves GPU utilization by allowing multiple users or applications to run on a single GPU simultaneously, each in its own isolated environment. However, current simulation tools fall short in modeling these scenarios, limiting our ability to predict and optimize the performance of multi-instance devices.

This thesis aims to extend existing simulation tools to accurately model the behavior and performance of multi-instance GPU devices. By integrating the unique characteristics of MIG technology into these tools, the research will provide insights into how workloads can be distributed across instances, how isolation impacts performance, and how resources are effectively shared. The ultimate goal is to enable more efficient use of multi-instance GPUs, driving better resource allocation strategies and advancing the field of GPU workload modeling.

**PREREQUISITES:** C/C++, Familiarity with GPU Architectures and Linux

**RELATED MATERIAL:** https://arxiv.org/abs/2409.06646, https://www.nvidia.com/en-eu/technologies/multi-instance-gpu/, https://doi.org/10.1109/ISCA45697.2020.00047

**CONTACT INFORMATION:**

Asst. Prof. Sotirios Xydis (sxydis@microlab.ntua.gr)

Konstantinos Iliakis (kiliakis@microlab.ntua.gr)

Panagiotis Eleftherakis (pelef@microlab.ntua.gr)